Reusing data produced by scientists

Karima Rafes
BorderCloud & INRIA

karima.rafes@gmail.com - @karima_rafes - http://karimarafes.me

Thesis supervisors: **Serge Abiteboul,** INRIA, ENS Cachan **Cécile Germain,** Université Paris-Sud, LRI
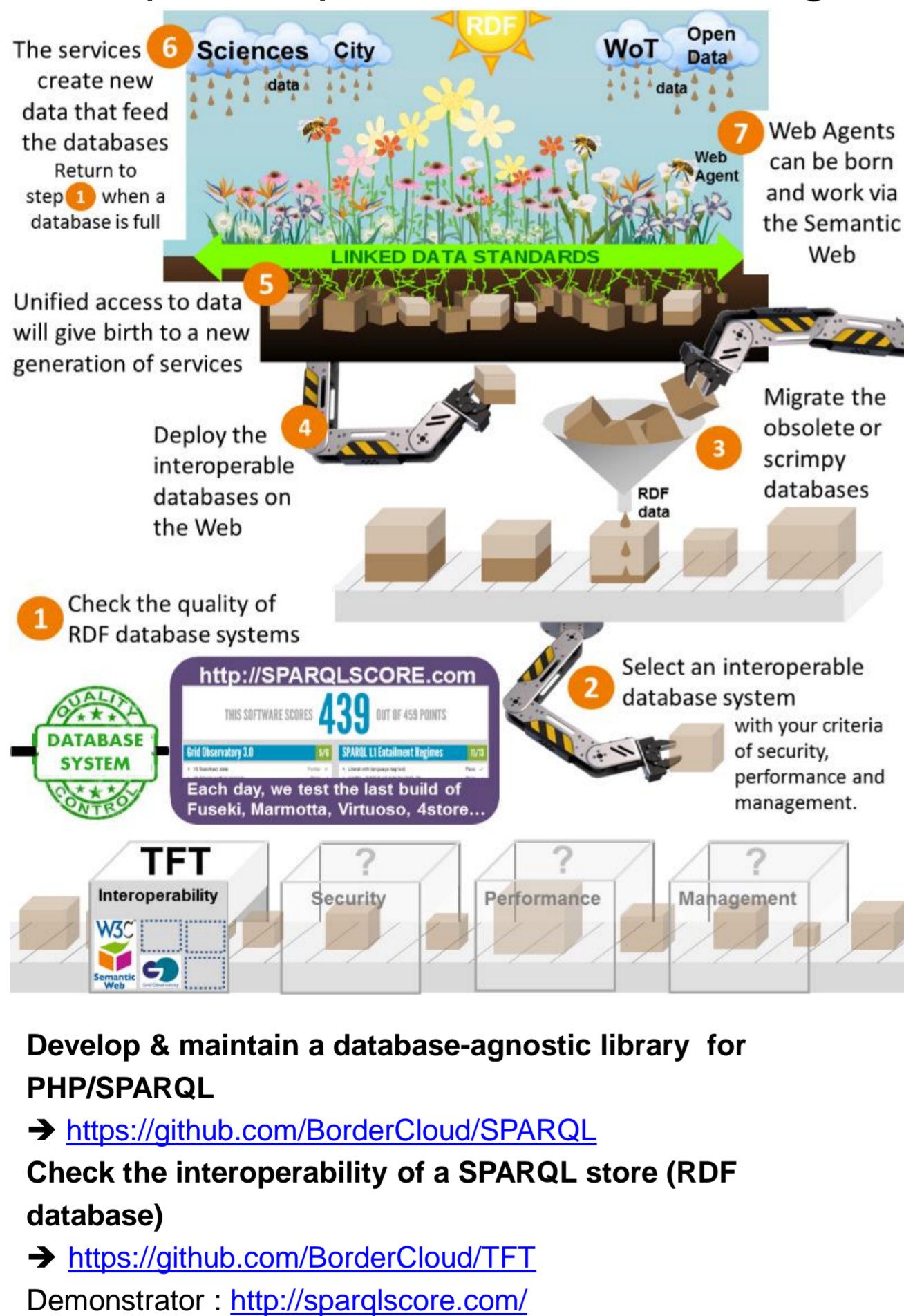
Thesis topic
Scientific data management using Semantic Web

1. Help them publish their knowledge
2. Construct/modify their knowledge
3. Find and reuse datasets of other scientist

Platform as a Service
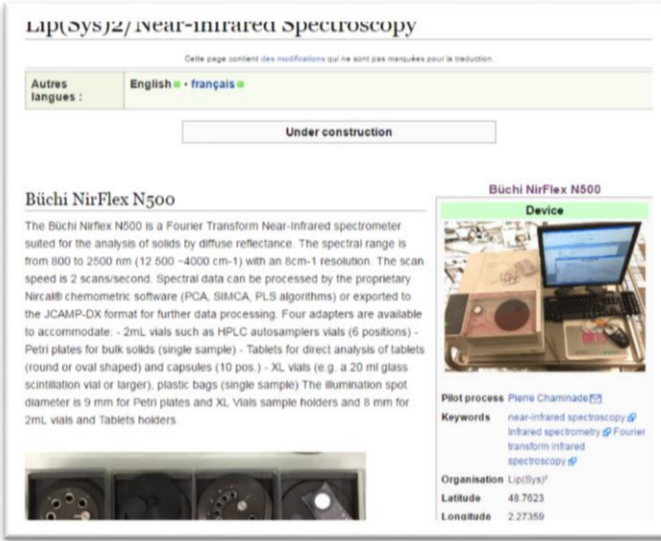
Help them publish their knowledge

The services create new data that feed the databases Return to step ① when a database is full

Unified access to data will give birth to a new generation of services

⑥ Sciences City data   WoT   Open Data data

⑦ Web Agents can be born and work via the Semantic Web

Migrate the obsolete or scrimpy databases

① Check the quality of RDF database systems

② Select an interoperable database system with your criteria of security, performance and management.

④ Deploy the interoperable databases on the Web

http://SPARQLSCORE.com
439 OUT OF 416 POINTS

Each day, we test the last build of Fuseki, Marmotta, Virtuoso, 4store...

Develop & maintain a database-agnostic library for PHP/SPARQL
➔ https://github.com/BorderCloud/SPARQL

Check the interoperability of a SPARQL store (RDF database)
➔ https://github.com/BorderCloud/TFT

Demonstrator : http://sparqlscore.com/

Knowledge base management
Construct their knowledge

Public knowledge bases construction

Project: DAAP
Data Acquisition For Analytical Platform
http://daap.eu/

http://daap.eu/wiki/Lip(Sys)2/Near-infrared_Spectroscopy

An extension for Mediawiki to:
- Record the metadata via the Wiki in a **synchronous** / **asynchronous** manner depending of the size of the team
- Display the metadata directly in the Wiki
- Compare to other ontologies via infoboxes

Project: Gregorius 3.0
Data about canon law
http://gregorius.fr/

http://gregorius.fr/wiki/Adrien_IV

Red flags marks the difference with Wikidata and their ontology. Scientists can detect the error simply in Wikipedia or their wiki.

Private knowledge base construction
Project: DAAP
Attach metadata to experimental data from the workflow in the lab

1. Write metadata in their private Wiki and generate the filenames

2. A tool reuse the metadata and simplify the workflow of their data (archive in a SAN)

3. In Matlab, scientists select their files in the cloud of the university via a SPARQL query on the private metadata.

Explore the new use cases with these new data

For example in Wikipedia.
https://io.datascience-paris-saclay.fr/appDisplayINRIATeams.php

These new knowledge bases will be very useful in order to discover for a topic of interest:
- Public datasets
- Available devices
- Other teams
- Experts

Interface for dataset reuse
Find and reuse datasets
https://io.datascience-paris-saclay.fr

Use case : Producer of data

1. Share a dataset

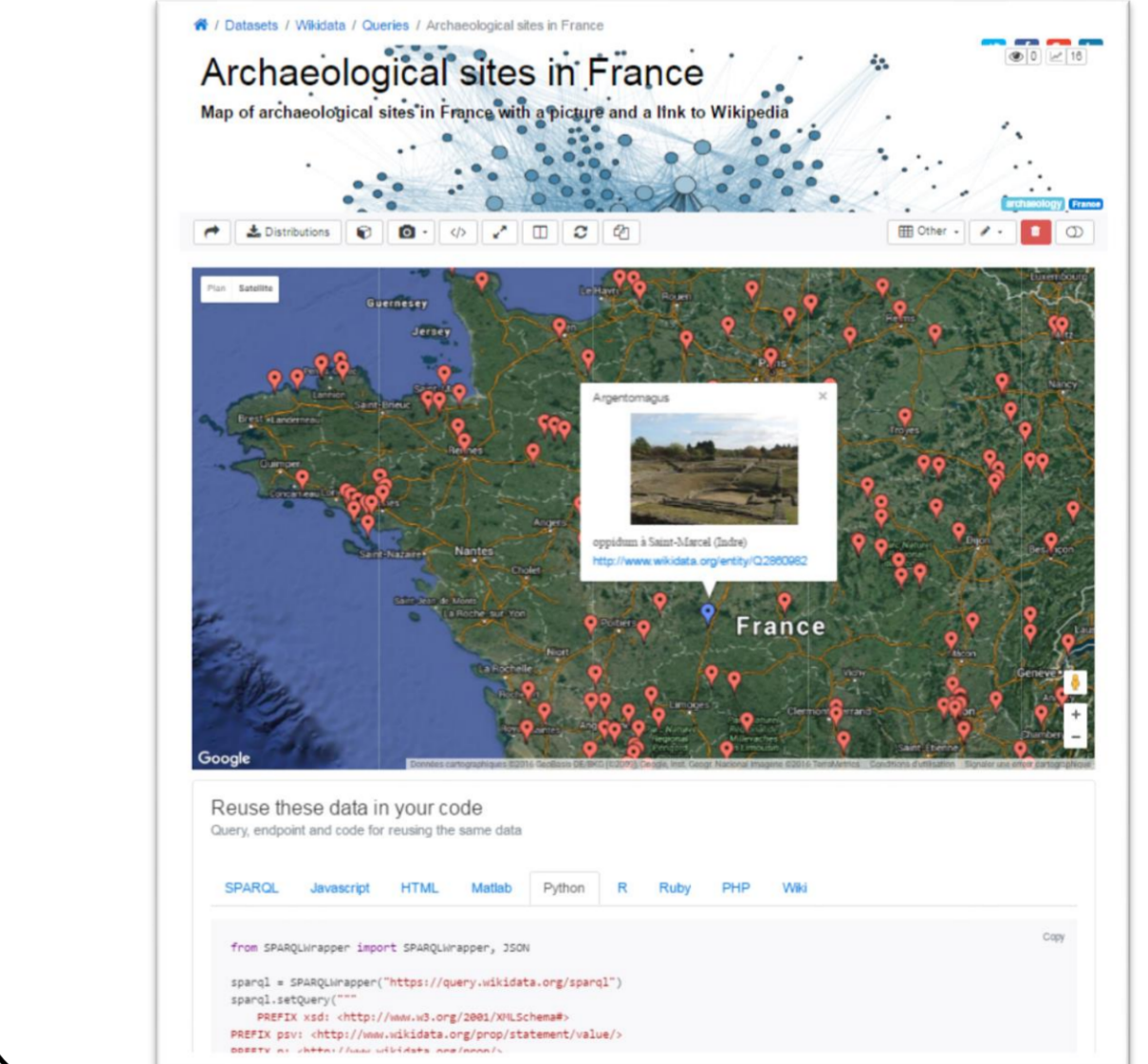**2. Build and share SPARQL examples in this mother tongue**

Use case : Consumer of data

1. Discover the datasets via Google & Wikipedia

2. Choose/copy/paste examples of queries in their code

Conclusion

Work towards a comprehensive open-source platform for scientists
- Several other projects on the same topics
- particularity: extensive use of RDF and linked data technology

Future work
- Generate new datasets via federated queries and export them in the cloud of the university.
- Optimize the scientists' Wikis for their measures' metadata.
- Deploy a open version of service for developers SPARQL in the world.

Demo : Test for triplestore

Demo : Interface for dataset reuse